PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2008) **275**, 871–878
doi:10.1098/rspb.2007.1558
Published online 15 January 2008

The economics of altruistic punishment and the maintenance of cooperation

Martijn Egas[1,*] **and Arno Riedl**[2]

[1]*Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, PO Box 94084, 1090 GB Amsterdam, The Netherlands*
[2]*Department of Economics, Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands*

Explaining the evolution and maintenance of cooperation among unrelated individuals is one of the fundamental problems in biology and the social sciences. Recent findings suggest that altruistic punishment is an important mechanism maintaining cooperation among humans. We experimentally explore the boundaries of altruistic punishment to maintain cooperation by varying both the cost and the impact of punishment, using an exceptionally extensive subject pool. Our results show that cooperation is only maintained if conditions for altruistic punishment are relatively favourable: low cost for the punisher *and* high impact on the punished. Our results indicate that punishment is strongly governed by its cost-to-impact ratio and that its effect on cooperation can be pinned down to one single variable: the threshold level of free-riding that goes unpunished. Additionally, actual pay-offs are the lowest when altruistic punishment maintains cooperation, because the pay-off destroyed through punishment exceeds the gains from increased cooperation. Our results are consistent with the interpretation that punishment decisions come from an amalgam of emotional response and cognitive cost–impact analysis and suggest that altruistic punishment alone can hardly maintain cooperation under multi-level natural selection. Uncovering the workings of altruistic punishment as has been done here is important because it helps predicting under which conditions altruistic punishment is expected to maintain cooperation.

Keywords: altruistic punishment; cooperation; Internet experiment; public good

Abbreviations: EMU; experimental money unit; ESM; electronic supplementary material; PP; punishment point

1. INTRODUCTION

Research has identified a variety of conditions necessary to sustain the evolution of cooperation. Such conditions include genetic relatedness among cooperators (West *et al.* 2002; Sachs *et al.* 2004), situations allowing for direct benefits to the cooperator (Sachs *et al.* 2004; Lehmann & Keller 2006), and repeated interactions allowing for indirect benefits to the cooperator via reciprocal altruism and reputation building (Trivers 1971; Axelrod & Hamilton 1981; Nowak & Sigmund 2005).

Recently, altruistic punishment has been proposed as a new mechanism maintaining cooperation in humans in the absence of any of the above-mentioned conditions. In behavioural experiments, altruistic punishment has been shown to effectively enforce cooperation among unrelated and anonymous humans (Fehr & Gächter 2002; Fehr & Fischbacher 2003). In one-shot interactions, people punish uncooperative behaviour at a cost to themselves, inducing future cooperation of the sanctioned individuals, in the absence of direct or indirect benefits to the punisher. These results challenge our view of human behaviour in social dilemma situations—cooperation seems to be maintained even in the absence of traditional mechanisms such as reciprocity and reputation. Importantly, in the mentioned experiments, the conditions for altruistic punishment were relatively favourable, however: low cost for the punisher and high impact on the punished.

Theory shows that free-riding and punishment-enforced cooperation are alternative stable states in simplified versions of the altruistic punishment game (Sigmund *et al.* 2001; Boyd *et al.* 2003). This bistability occurs because rare punishers in a group of free-riders would lose out since they incur high costs from punishing all free-riders, and rare free-riders in a group of punishers would lose out because they experience lots of punishment. Hence, a critical mass of punishers is needed to guarantee effective sanctioning of free-riding. Therefore, if the number of altruistic punishers or the amount of altruistic punishment significantly responds to variations in costs and impact of punishment, it can have dramatic effects on cooperation. Indeed, recent experimental studies suggest that the punishment costs an individual incurs affect non-altruistic punishment behaviour negatively (Anderson & Putterman 2006; Kosfeld & Riedl 2007; Carpenter 2007; Nikiforakis & Normann in press). However, it is still largely unknown whether and under what conditions different impact-to-cost ratios (hereafter called effectiveness) can lead the dynamics of altruistic punishment and cooperation to the different alternative states of cooperation and free-riding.

The aim of our research is to identify the boundaries of altruistic punishment for maintaining cooperation among

*Author for correspondence (egas@science.uva.nl).

All authors contributed equally to this work.

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2007.1558 or via http://journals.royalsociety.org.

Received 12 November 2006
Accepted 12 December 2007

unrelated individuals. To this end, we systematically varied both the cost and the impact of altruistic punishment. We used the Internet to perform a large-scale experiment, thereby obtaining a large and relatively heterogeneous subject pool. Our results show that the effectiveness of altruistic punishment is an important determinant of punishment behaviour, and thereby a crucial parameter for the maintenance of cooperation. Importantly, altruistic punishment takes place for all the investigated levels of effectiveness. However, the threshold level at which free-riding behaviour goes unpunished is strongly increasing with decreasing effectiveness of punishment. Interestingly, the behaviour of participants from our extensive subject pool is qualitatively comparable with that of students. This observation is in line with the results of other recent experiments with non-standard subject pools investigating punishment behaviour in various small-scale societies (Henrich *et al.* 2006). Our finding that the success of altruistic punishment seems to depend on only one variable—the threshold level of free-riding that goes unpunished—can be valuable for theoretical model building and help to explore the circumstances under which altruistic punishment is likely to maintain cooperation.

2. MATERIAL AND METHODS

(a) *The public goods game with punishment*

We conducted an altruistic punishment experiment with real money at stake where we systematically varied both the cost and the impact of punishment (see §2*b* for details). We implemented four treatments with punishment and a control treatment without punishment. In each treatment, subjects were engaged in a public good game in groups of three. Each subject was endowed with 20 experimental money units (EMUs) and could contribute between 0 and 20 EMU of this endowment to a group project. In the implemented setting, material self-interest dictates to contribute nothing to the group project, whereas collectively it is optimal for everybody to contribute the entire endowment. After decisions in the public good game were made, each group member was informed about the other group members' contributions and the resulting earnings. In the treatment without punishment, this ended the interaction between the group members. In the punishment treatments, each member had the possibility to punish other members by assigning between 0 and 10 punishment points (PPs) to each of the two other members. Importantly, the punishment treatments differed in the cost and the impact of punishment. In treatment T13, which is akin to the standard altruistic punishment experiment (Fehr & Gächter 2002), each assigned PP costs the punisher 1 EMU and reduces the pay-off of the punished with 3 EMU. In T31, the costs per assigned PP were 3 EMU for the punisher and only 1 EMU for the punished. In treatments T11 and T33, this relation was 1 : 1 and 3 : 3, respectively. For convenience, we call this impact-to-cost ratio *effectiveness* of punishment. The control treatment without punishment is indicated by T00. The design guaranteed full anonymity of subjects. Because the act of punishing comes at a cost, purely selfish subjects will never punish. Given that nobody punishes and because not contributing is a dominant strategy in the public good game, selfish individuals will also not contribute in any treatment.

To allow for learning, the experiment was repeated for six rounds in each replicated group of subjects. To exclude the potential effects of reciprocation (Trivers 1971; Axelrod & Hamilton 1981) and reputation building (Sugden 1986; Alexander 1987; Nowak & Sigmund 2005), we implemented a so-called 'perfect stranger' design which ensured that no subject ever met another subject more than once. Hence, also in the repeated interaction, purely selfish individuals neither punish nor contribute to the public good. Previous work (Fehr & Gächter 2000, 2002), however, indicates that at least in environments where punishment is relatively cheap and has strong impact, the frequency of altruistic punishment is surprisingly high, stimulating high contribution rates in the public good game. Below, we show that these results do not survive our variations in the effectiveness of punishment.

We implemented a large-scale set-up with in total 846 participants and used the Internet to facilitate this experiment. In our experiment, any Dutch-speaking person could participate. This set-up has two important implications. First, it extends our subject pool beyond the typically used undergraduate students and allows us to assess the robustness of results from such laboratory experiments. The socio-economic characteristics of the participants confirm that our participants differ from a pool of students. The average gross income of our subjects was close to the actual average gross income in The Netherlands. The average age was 35 years (range: 12–80 years), and education ranged from secondary school (3%) up to university degrees (33%). Female participants (28%) were under-represented. (See the electronic supplementary material for details of the subject pool characteristics and the recruitment method. There are some potential pitfalls in using the Internet for experimental games. We discuss these and the way we responded to them with our experimental procedures at length in the electronic supplementary material.)

Second, some scholars argue that the narrow subject pool (e.g. students from the same college as in Fehr & Gächter (2002) and people from the same village as in Henrich *et al.* (2006)) usually used in experiments may (unconsciously) trigger a group identity. This may undercut the explicit experimental features of anonymity and one-shot encounters because subjects may (unconsciously) perceive the situation as non-one shot and non-anonymous (Hagen & Hammerstein 2006). Our large-scale set-up and the obvious anonymity of interactions via the Internet circumvent or at least drastically reduce this potential problem. Our results (below) show that in the standard treatment the behaviour of our participants does not differ qualitatively from the behaviour of student participants in similar experiments (Fehr & Gächter 2000, 2002; Fehr & Fischbacher 2003). Other research finds similar results for experiments on altruistic punishment with subject pools consisting of students in Germany, Switzerland, Russia and Belarus, as well as rural and urban non-students in Russia (Gächter & Herrmann 2006; S. Gächter 2006, personal communication). This lends confidence to the robustness of these earlier experimental results and indicates that the use of a narrower subject pool does not necessarily impair the obtained results.

When investigating punishment behaviour, we focus on the difference in contributions between the punisher and the punished individual. Compared with another widely used measure, the deviation from the group average, our measure has two advantages. First, this is the most salient measure for triggering punishment, because it approximates how much

one individual free-rides on the contributions of the other individual. Second, the idea that the group average should be taken as contribution norm is not necessarily shared by all group members and needs a degree of coordination that is hardly achievable in our anonymous setting and the perfect stranger design. For convenience, we focus our attention on cases where the punished person contributed less than (or the same as) the punishing person. Some low level of punishment does occur when co-participants invest more, which is taken into account in our analysis whenever (statistically) necessary (i.e. tables S1–S4, S6; see table S1 in the electronic supplementary material for more details on such so-called 'counter-intuitive punishment').

(b) *Experimental procedures*

In our experimental set-up of the public good game, each EMU invested returned 0.5 EMU to each of the three group members. From an individual perspective, each EMU kept paid off one EMU, whereas each invested EMU only paid off 0.5 EMU. Hence, material self-interest dictated to contribute nothing to the group project. If all group members kept their endowment, everybody earned 20 EMU. If all contributed their entire endowment to the group project, then each would earn $60 \times 0.5 = 30$ EMU. Decisions were made simultaneously and anonymously. One replicated group consisted of 18 subjects, exposed to exactly one treatment.

A total of 846 people participated in the experiment. The experiment was conducted via the Internet on a secure server using client-based software. The subject pool consisted of volunteers from the Dutch-speaking (world) population with Internet connection. The recruitment took place via mass media in The Netherlands (advertisements in newspapers and on radio) and the science website of the Dutch public broadcasting station VPRO. In all recruitment announcements, we made sure that the actual content of the experiment was not revealed. The only information about the experiment given during the recruitment period was that a scientific experiment will take place with the possibility to earn money. (The title of the experiment was 'Speel je rijk', which loosely translates as 'Play to get rich'.) No further information about the content was revealed until the last experimental session was finished. Furthermore, it was announced that the experiment was going to take place from 24 to 28 May 2004, with two sessions per day (at 16.00 and 20.30): only one person is allowed to participate in one session, and that a session will take approximately 45–60 min. A person interested in participating was asked to send an e-mail and to indicate two preferred sessions (dates and times). They were then sent an acknowledgement e-mail. This e-mail contained the following information: (i) a random lottery will decide whether (s)he is chosen as an actual participant and (ii) if (s)he is chosen, this information will be transmitted shortly (usually 24 hours) before the chosen session takes place. All together more than 4000 people subscribed for the experiment. From these, approximately 1000 were randomly selected as participants, 846 of which actually participated (not all selected people 'showed up' at the experiment).

A session was organized as follows. All participants received an e-mail with a password and the website address from where the experiment was going to start. With the password and his or her e-mail address, a participant could log into the experiment. There each participant received online instructions that explained the structure of the experiment in detail. That is, it was explained to the participants how to make decisions, how to calculate earnings, how their own earnings and the earnings of others depend on their own decisions and the decisions of others; that they were going to play six rounds in groups of three; that these groups are going to be recomposed after each round to guarantee that nobody meets anyone twice; that all interactions will be anonymous; and that the history of the behaviour of participants was not going to be disclosed to anybody. After having read the instructions, each participant had to answer a number of control questions which allowed us to check that (s)he understood the instructions. (Experimental instructions are available upon request from the authors.) These questions concerned the reshuffling of the groups after each round, the consequences of (not) contributing and (not) punishing, and the calculation of one's own earnings and the earnings of other group members in a number of hypothetical situations. Only if the subject answered all questions correctly (s)he was allowed to participate in the experiment. (Only a few subjects dropped out in this phase.) During the instructions and the control questions, the subjects had the possibility to ask questions to the experimenters (Martijn Egas and Arno Riedl) using a chat window that was built in the software.

When a subject had answered all the questions correctly, (s)he entered a waiting queue until a group of 18 participants was formed. Each of the 18 participants then played six rounds of the public good game with or without punishment, depending on the treatment. The timing of our five treatments (T00, T13, T11, T31 and T33) was determined beforehand and guaranteed a balanced distribution of afternoon and evening sessions across treatments. In each experimental session, we implemented only one treatment. Since the subjects could only participate in one session, each participant was only participating in one treatment. Subjects were not aware of the fact that there were different treatments. The number of subjects participating in a session varied between 54 (three groups of 18) and 126 (seven groups of 18). In total, 10 replicate groups of 18 participated in T13, T11 and T31; 9 in T33 and 8 in T00.

In the first round of a session, the 18 participants of a group were randomly allocated to six subgroups of three. In the five subsequent rounds, the groups of three were recomposed such that nobody met anybody else twice. No group member knew anything about the past behaviour (contribution to the public good and punishment decisions) of the other group members. In each round, everybody made a contribution decision simultaneously, which was subsequently disclosed to the other two subgroup members. Also, the earnings of all three subgroup members were shown to each of the subgroup members. Hence, everybody knew, in principle, the earnings of the other members in the subgroup. In T00, this ended the round, but in the treatments with punishment this was the stage where participants could assign between 0 and 10 PPs (called 'deduction points' in the experiment) to each of the other two subgroup members. Finally, each participant was informed how many PPs in total (s)he received and what the net earnings over this round amounted to. Total earnings were accumulated across rounds.

After the sixth round, participants were asked to fill out a questionnaire with questions on their socio-economic background. The participants were informed that they had to answer all questions in order to be able to receive the money they earned in the experiment. Since some of the questions could be regarded as sensitive and/or private information, the

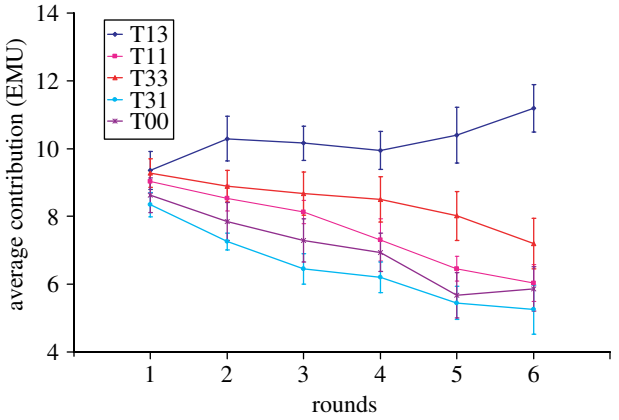

Figure 1. Changes over rounds in the average (±1 s.e.m.) contribution to the public good. In all five treatments, the initial contribution rate is approximately 9 EMU (45% of the endowment). Only in T13, cooperation increases over the six rounds of the experiment; in all other cases, cooperation is in clear decline (repeated-measures ANOVA, treatment: $F_{4,42}=11.522$, $p \ll 0.001$; round: $F_{5,210}=18.146$, $p \ll 0.001$; treatment×round: $F_{20,42}=4.056$, $p \ll 0.001$). Tukey's *post hoc* tests showed that contributions in T13 differed from all other treatments, and that T33 differed from T31. Blue diamonds, T13; pink squares, T11; red triangles, T33; turquoise circles, T31; purple asterisk, T00.

option 'no answer' was provided for each question (except age). The subjects were informed about this before they entered the questionnaire. Complete anonymity as well as privacy protection was guaranteed because the answers are only used by the experimenters for scientific ends, and not coupled to the names of the participants. The participants were informed about this. All participants except three filled in the questionnaire. Furthermore, an overwhelming majority did not use the option no answer for any question, although use of this option varied with the content of the questions (see 'Socio-economic characteristics of the subject pool' in the electronic supplementary material).

Finally, each participant was asked to give his or her name, place of residence and bank account number to be able to transfer their earnings. A group of 18 participants was typically finished after 40 min. The average earnings were 12.20 euro per participant.

3. RESULTS

Our first result concerns the degree of cooperation in the public good game. If punishment is anticipated, one may expect differences in initial contribution rates between the punishment treatments and the control treatment (see Fehr & Gächter 2000, 2002). We find no evidence for this (similar to Walker & Halloran (2004) and Gächter & Herrmann (2006)). In all treatments, average contributions start off around the same level (figure 1; Kruskal–Wallis test: $\chi^2_4=3.042$, $p=0.551$ on group-level data; $\chi^2_4=3.167$, $p=0.530$ on individual-level data). However, the dynamics of cooperative behaviour over the six rounds are strikingly different across treatments. Only when punishment is relatively effective (T13), contributions increase over rounds. In all other treatments, however, contributions are quickly declining (figure 1). This shows that the scope for punishment to maintain cooperation in the long run is clearly dependent on both its cost (comparison T13 and T33) and its impact (comparison T13 and T11).

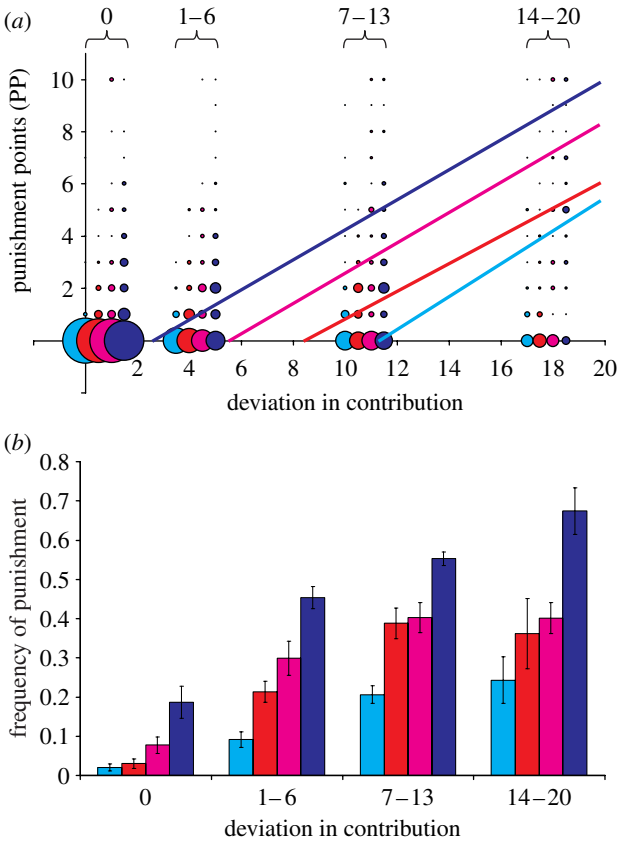

Figure 2. Punishment characteristics as a function of the deviation in contribution. (*a*) Using four categories (0, 1–6, 7–13 and 14–20) of the deviation in contribution between the punisher and the punished participant, the size of each circle indicates the relative frequency of PPs allocated to participants with such deviations in contribution (data are based on all six rounds). Straight lines represent linear Tobit regressions for each of the four punishment treatments. PPs dealt out are significantly increasing with deviation in contribution. Slopes of the regression lines are not significantly different from each other. Estimated deviation thresholds (TH) up to which deviation in contribution goes unpunished are significantly different and have the order $TH(\text{T13}) < TH(\text{T11}) < TH(\text{T33}) < TH(31)$. Statistical details in the electronic supplementary material, table S1. Turquoise circles, T31; red circles, T33; pink circles, T11; blue circles, T13. (*b*) Frequency of punishment of defecting participants (data are based on all six rounds). Logit regression analysis shows that the marginal likelihood that deviating participants are punished is significantly increasing with increasing deviation in contribution in all treatments. These marginal propensities to punish are the same in all treatments. Statistical details in the electronic supplementary material, table S2. Turquoise bars, T31; red bars, T33; pink bars, T11; blue bars, T13.

Note that this result is in line with the above-discussed theoretical prediction that cooperation—induced by (the threat of) punishment—and defection are alternative attracting states.

In all punishment treatments, inflicted punishment strongly depends on the difference in contributions between the punisher and the punished. Generally, punishers allocate increasingly more PPs the more the other's contribution falls short of their own contribution (figure 2*a*). For convenience, in the following we use 'deviation in contribution' as the difference between the contribution of the focal participant and the contribution of her co-participant.

The systematic variation of cost and impact of punishment allows us to investigate how participants change behaviour in response to changes in these critical parameters. Comparing punishment behaviour across treatments clearly shows that significantly more PPs are dealt out when it is cheap and has high impact (T13) than when it is expensive and has low impact (T31). For intermediate effectiveness, low cost and low impact (T11) and high cost and high impact (T33), the allocated PPs lie between the other two treatments. To examine the differences between treatments statistically, we performed Tobit regressions with PPs as dependent variables and (negative and positive) deviations in contribution as independent variables (figure 2*a*; table S1 in the electronic supplementary material). Surprisingly, the marginal propensity to increase punishment with increasing deviation in contribution is the same for all four punishment treatments. This is manifested by the equality of the regression coefficients of deviations in contribution in all treatments (for details see figure 2*a* and table S1 in the electronic supplementary material). However, the cost and the impact of punishment have a significant effect on the *threshold* of deviation in contribution at which participants start to punish free-riders. This threshold is significantly increasing with decreasing effectiveness from 2.41 (T13) to 5.34 (T11) to 8.33 (T33) to 11.3 (T31; two-sided nonlinear Wald tests, $p<0.03$ in all cases).

For the frequency of punishing free-riders, equivalent results are found. Examination based on logit regressions (table S2 in the electronic supplementary material) shows that the differences in punishment frequencies are solely due to a shift in the deviation threshold at which participants start to punish free-riders. The marginal propensity to punish is the same in all treatments. Furthermore, the frequency of punishment is monotonically increasing with the deviation from the punishers' contributions (figure 2*b*; table S2 in the electronic supplementary material).

The monetary effect of punishment for the punished (i.e. the average amount of EMU lost due to received punishment, as a function of deviation in contribution) differs rather dramatically between the most effective treatment T13 and the other three punishment treatments. In all the categories of deviation in contribution, punished participants suffer much more in T13 than in the other treatments. Tobit regressions show that in T13 already very small deviations in contribution (1.43) lead to a noticeable effect of punishment (figure 3; table S3 in the electronic supplementary material). In the other treatments, the punishment effect sets in only at large deviations (6.83 in T11, 7.35 in T33 and 12.0 in T31). Note that the effect of punishment in EMU is remarkably similar in treatments T11 and T33, as may be expected when punishment effectiveness governs this variable, rather than any other interactive effect of impact and cost of punishment. An equivalent result is found when analysing expenditures for punishment (figure S2 and table S4 in the electronic supplementary material). One surprising upshot of these results is that the force of punishment effectiveness can be pinned down to one single variable: the threshold level of free-riding that goes unpunished. The shift in this variable alone suffices to explain whether cooperation is sustained or deteriorates (figure 1).

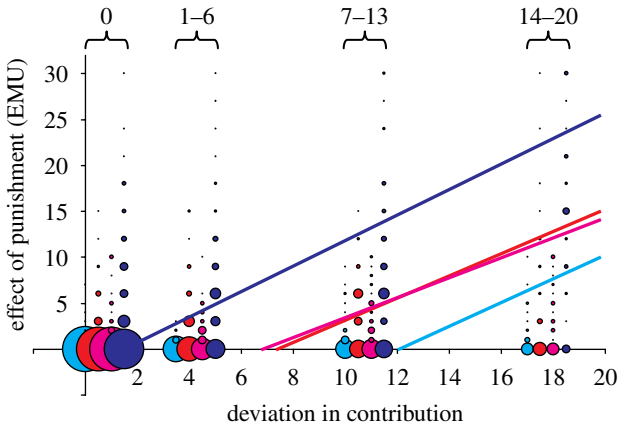

Figure 3. The monetary effect of punishment on punished participants. Using four categories (0, 1–6, 7–13 and 14–20) of the deviation in contribution, the size of each circle indicates the relative frequency of the monetary effect of punishment (in EMU) on participants with such deviations in contribution (data are based on all six rounds). Straight lines represent linear Tobit regressions for each of the four punishment treatments. Slopes of the regression lines are not significantly different from each other except for the comparison of T13 with T11. Estimated deviation thresholds (*TH*) up to which deviation in contribution goes unpunished are significantly different and have the order $TH(\text{T13}) < TH(\text{T11}) = TH(\text{T33}) < TH(31)$. Statistical details in table S3 in the electronic supplementary material. Turquoise circles, T31; red circles, T33; pink circles, T11; blue circles, T13.

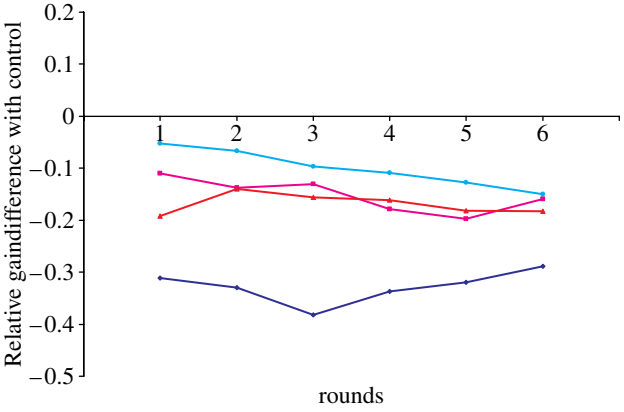

Figure 4. Altruistic punishment does not pay off. The average relative gains ((earnings in punishment treatment-earnings in no-punishment treatment)/earnings in no-punishment treatment) neither significantly increase nor decrease over the six rounds of the experiment. (Spearman rank-order correlation; T13: $\rho=0.7143$, $n=6$, $p=0.1108$; T11: $\rho=-0.0857$, $n=6$, $p=0.8717$; T33: $\rho=0.0857$, $n=6$, $p=0.8717$; T31: $\rho=-0.7714$, $n=6$, $p=0.0724$). Relative gains in round 6 are all significantly smaller than zero ($p<0.02$, two-sided *t*-tests). Blue diamonds, T13; pink squares, T11; red triangles, T33; turquoise circles, T31.

Considering the actual pay-off per group uncovers a sobering picture. In one treatment (T13) where punishment increases cooperation, groups earn significantly less than groups in any of the other treatments (figure 4). Furthermore, compared with the control treatment without punishment, earnings in T13 (as well as all other punishment treatments) are clearly inferior and show no statistically significant tendency to catch up (figure 4). Moreover, the average actual earnings in T13 are even

lower than the potential earnings of full free-riding without punishment. The reason for this result is that the opportunity to punish is used most frequently in T13, but the increase in contributions due to this punishment is not sufficient to compensate for the cost of punishment. Hence, in our experiment, altruistic punishment leads to an overall loss of individual and group welfare. It might be expected that in the long run actual earnings in T13 would rise and eventually exceed that in the other treatments as a result of increasing cooperation and decreasing incidence of punishment. However, our results indicate that it will probably take many more rounds before the cumulative income in T13 may exceed that in the other treatments.

A distinguishing characteristic of our subject pool is its heterogeneity in the socio-economic backgrounds of the participants. We can exploit this fact and identify socio-economic determinants of contribution and punishment behaviour. We conducted regression analyses with contributions to the public good and allocated PPs, respectively, as dependent variables. In both the cases, we examined two models. In model 1, only age and sex are used as socio-economic explanatory variables. In model 2, a number of other socio-economic background variables are added (for details see 'Socio-economic determinates of contribution and punishment behaviour' in the electronic supplementary material). Contributions to the public good seem weakly positively influenced by age ($p=0.051$ and 0.805 in models 1 and 2, respectively) and are independent of participants' sex ($p>0.2$ in both models). Interestingly, our model 2 strongly indicates that being a young student (or pupil) has a strong and significantly negative effect on contributions to the public good (regression coefficient $=-5.516$, $p=0.023$; interaction with age: coefficient $=0.218$, $p=0.025$). All other investigated socio-economic variables have no significant effect on contributions. The investigation of punishment behaviour shows that being older and being male significantly increases the amount of allocated PPs ($p<0.037$ for age and $p<0.051$ for being male, in both models). Interestingly, of the other investigated variables only being a student shows a (marginally) significant and negative effect ($p=0.085$), indicating that students are punishing less, *ceteris paribus*.

4. DISCUSSION

The observed behaviour in treatment T13 could be explained by proximate fairness models (e.g. Fehr & Schmidt 1999; Fowler *et al.* 2005) assuming that inequity-averse individuals will punish free-riders because punishing reduces pay-off differences between punishers and punished. In treatments T11, T33 and T31, however, such a model cannot explain the still existing non-negligible amount of punishment. In these treatments, punishment either leaves the relative pay-offs the same or even increases the inequality to the disadvantage of the punisher. Under such conditions, fairness theory based solely on outcomes predicts behaviour that is indistinguishable from pure selfishness. To explain punishment in these treatments, one has to resort to proximate fairness models that take reciprocal inclinations and intentions more directly into account (Rabin 1993; Levine 1998; Dufwenberg & Kirchsteiger 2004; Falk & Fischbacher 2006). It is still an open question what the sources of altruistic reciprocal behaviour, i.e. punishment in our experiment, are. Recent research has identified emotions as one possible source (Bosman & Van Winden 2002; Sanfey *et al.* 2003; de Quervain *et al.* 2004; S. Gächter, personal communication, but see also Knoch *et al.* (2006) who challenge this view) and our results are consistent with such an interpretation. However, the strong response in punishment behaviour to our variations in the cost and the impact of punishment indicate that the decision to altruistically punish is also strongly influenced by material economic incentives. Taken together, this suggests that the 'decision to punish' comes from an amalgam of a non-optimizing (in a material sense), probably emotional, response and cognitive material cost–impact analysis, a view supported by the recent neuro-economical findings (de Quervain *et al.* 2004; Knoch *et al.* 2006).

Given the strong economic relationships, punishment appears to require both low cost and high impact to maintain cooperation. Note, however, that under these conditions the signalling value of altruistic punishment to increase cooperation may be under threat. In situations where individuals do not share a strong common interest, signals are expected to remain honest only when they are costly (Maynard Smith & Harper 2003), i.e. low-cost punishment may be abused by defectors to induce higher cooperation in cooperators, which would undermine the effect of punishment on cooperation. Reputation systems implemented by electronic trading platforms such as eBay may serve as an example: it allows cheap punishment of dishonest sellers by negative reputation feedback, yet does not improve sellers' performance, possibly because the published reputation carries little truthful information (Resnick & Zeckhauser 2002). Indeed, experiments show that feedback reputation mechanisms are inferior to direct interactions for developing trust (Bolton *et al.* 2004).

Our results show that the total pay-off in a group is the lowest when altruistic punishment successfully enhances cooperation. Interestingly, the data gathered in the experimental study most akin to our experiment (Fehr & Gächter 2002; experimental sequence 1) show a very similar pattern. The result is not reported in Fehr & Gächter (2002) but our own calculations using their raw data show that average relative gains in the punishment treatment vary between -0.25 and -0.08. Hence, as in our experiment, total earnings in the punishment treatment are always below those in the control treatment without punishment. The relative gains are increasing first, flatten out from round 4 onwards and even decrease from the penultimate round (-0.08) to the last round (-0.10). The reason is that the maintenance of cooperation requires continued active punishment to such an extent that the costs do not outweigh the benefits of increased co-operation. Although one should not take a one-to-one relation between material pay-offs and utility or fitness for granted, we believe that in our context the material pay-offs are a good if imperfect proxy of actual well-being. In this sense, our result sheds some doubt on group selection models for the evolution of cooperation by altruistic punishment, where sanctioning within a group is supposed to increase cooperation and thereby the competitive strength of that group (Boyd *et al.* 2003). It is a debated issue whether cooperation *per se* or the generated pay-off is crucial for the survival of groups in group contests. From a game theoretic viewpoint, pay-offs should determine a

group's competitive strength. However, cooperation *per se* might be a relevant factor because it could be hypothesized that, for example, altruistic punishers are not only sanctioning in-group deviations from cooperation but also inclined to reciprocate hostile acts by out-group members. Indeed, there is a preliminary evidence for some in-group favouritism in punishment of norm violations (Bernhard *et al.* 2006*a*). However, whether such behaviour can be generalized to the situations of group competition is, due to missing empirical evidence, an open question. Another critical issue of our (and many other) experimental studies is the relatively short time horizon. One might argue that with many more repetitions punished defectors eventually will learn to cooperate, and that at some point the mere threat of punishment will be sufficient for sustaining cooperation. In particular, in the light of the theoretical finding that free-riding and cooperation enforced through punishment are alternative stable states, it is unclear whether behaviour would converge to a state where the total gains are larger in groups with or in groups without punishment. Our evidence suggests at least two preconditions for altruistic punishment to be successful in material terms. First, the cost–impact ratio has to be relatively favourable for punishment, and second, it may take a very large number of repetitions, especially if the cumulative pay-offs are taken into account.

In experiments where participants are facing the same co-participants in every round (the so-called 'partner' set-up; Fehr & Gächter 2000, Masclet *et al.* 2003) cooperation levels rise faster, potentially allowing for higher earnings under effective punishment within a smaller number of rounds. Hence, in such an environment, it is clearly conceivable that multi-level selection may favour the enforcement of cooperation by punishment. However, since such a set-up allows for direct reciprocation and reputation building, such punishment cannot be considered 'altruistic punishment' (as defined in Fehr & Gächter 2002), but a form of 'costly punishment' where punishing may yield future material gains for the punisher (Brandt *et al.* 2006; Rockenbach & Milinski 2006).

Recent evidence shows that costly punishment can indeed be effective in maintaining cooperation in situations where reciprocity and reputation building are possible (Fehr & Gächter 2000; Masclet *et al.* 2003; Rockenbach & Milinski 2006) or when individuals can freely choose to implement such punishment rules (Gürerk *et al.* 2006). Also, the combination of costly punishment with reputation building through indirect reciprocations seems to be very effective in increasing the efficiency in public goods experiments (Rockenbach & Milinski 2006). Similarly, the (individual as well as collective) exclusion of free-riders from repeated interaction networks increases cooperation (Masclet 2003; Panchanathan & Boyd 2004; Ule 2005). In these examples, punishment also increases pay-offs, at least in some events. Casual evidence also suggests that 'gossiping' may be a cheap and effective way to punish free-riders. Note, however, that the effectiveness of such a mechanism relies on tight social groups or networks and effective means of communication, preconditions that are hardly met in a perfect stranger setting where people engage in strictly anonymous one-shot interactions.

Situations of repeated interaction, in dyads or in interaction networks, resemble some real-life settings more closely than a perfect stranger design does, although many encounters in everyday life are truly or almost one-shot interactions (e.g. helping strangers when travelling, tipping in non-hometown bars and restaurants, Internet shopping). Importantly, however, non-anonymous repeated interaction also generates opportunities that may undermine the effectiveness of costly punishment. For instance, such situations allow that individuals (threaten to) punish others who punish them. That such counter-punishment not only exists as a thought experiment is testified by modern real-life examples of 'vendetta' in, for example, Corsica, parts of Italy, northern Albania and eastern Turkey. Recent experimental evidence also shows that the possibility of counter-punishment has strong detrimental effects on cooperation as well as individual and group welfare (Nikiforakis 2008).

In conclusion, we find that altruistic punishment enforces cooperation only when its effectiveness is relatively high. Additionally, individual and group pay-offs are relatively low even if cooperation is successfully enforced. Other studies and real-life examples suggest that mechanisms involving repeated interactions, such as reciprocity, reputation, exclusion, parochialism (Bernhard *et al.* 2006*b*) and also opting out (Fowler 2005; Brandt *et al.* 2006) can have strong cooperation-enhancing effects. Taken together, the evidence indicates that altruistic punishment may be important in the evolution of cooperation only in combination with such other cooperation-enhancing mechanisms.

We thank Maarten Boerlijst, Ernst Fehr, Simon Gächter, Ellinor Ostrom, Ernesto Reuben, Maurice Sabelis and three anonymous reviewers for their helpful comments. We also thank Pieter van Beek, Laura van Geel and 'De Praktijk' for their research assistance. Administrative and financial assistance of VPRO/wetenschap in conducting the experiments is gratefully acknowledged. This paper is part of the research project 'Simultaneous evolution of social norms and social behaviour: a combined theoretical and experimental approach' funded by The Netherlands Organisation for Scientific Research (NWO).

REFERENCES

Alexander, R. D. 1987 *The biology of moral systems*. New York, NY: Aldine de Gruyter.

Anderson, C. M. & Putterman, L. 2006 Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. *Games Econ. Behav.* **54**, 1–24. (doi:10.1016/j.geb.2004.08.007)

Axelrod, R. & Hamilton, W. D. 1981 The evolution of cooperation. *Science* **211**, 1390–1396. (doi:10.1126/science.7466396)

Bernhard, H., Fehr, E. & Fischbacher, U. 2006*a* Group affiliation and altruistic norm enforcement. *Am. Econ. Rev. Papers Proc.* **96**, 217–221.

Bernhard, H., Fischbacher, U. & Fehr, E. 2006*b* Parochial altruism in humans. *Nature* **442**, 912–915. (doi:10.1038/nature04981)

Bolton, G. E., Katok, E. & Ockenfels, A. 2004 How effective are electronic reputation mechanisms? An experimental investigation. *Manage. Sci.* **50**, 1587–1602. (doi:10.1287/mnsc.1030.0199)

Bosman, R. & Van Winden, F. 2002 Emotional hazard in a power-to-take experiment. *Econ. J.* **112**, 147–169. (doi:10.1111/1468-0297.0j677)

Boyd, R., Gintis, H., Bowles, S. & Richerson, P. J. 2003 The evolution of altruistic punishment. *Proc. Natl Acad. Sci. USA* **100**, 3531–3535. (doi:10.1073/pnas.0630443100)

Brandt, H., Hauert, C. & Sigmund, K. 2006 Punishing and abstaining for public goods. *Proc. Natl Acad. Sci. USA* **103**, 495–497. (doi:10.1073/pnas.0507229103)

Carpenter, J. 2007 The demand for punishment. *J. Econ. Behav. Org.* **62**, 522–542. (doi:10.1016/j.jebo.2005.05.004)

de Quervain, D. J.-F., Fischbacher, U., Treyer, V., Schelhammer, M., Schnyder, U., Buck, A. & Fehr, E. 2004 The neural basis of altruistic punishment. *Science* **305**, 1254–1258. (doi:10.1126/science.1100735)

Dufwenberg, M. & Kirchsteiger, G. 2004 A theory of sequential reciprocity. *Games Econ. Behav.* **47**, 268–298. (doi:10.1016/j.geb.2003.06.003)

Falk, A. & Fischbacher, U. 2006 A theory of reciprocity. *Games Econ. Behav.* **54**, 293–315. (doi:10.1016/j.geb.2005.03.001)

Fehr, E. & Fischbacher, U. 2003 The nature of human altruism. *Nature* **425**, 785–791. (doi:10.1038/nature02043)

Fehr, E. & Gächter, S. 2000 Cooperation and punishment in public goods experiments. *Am. Econ. Rev.* **90**, 980–994.

Fehr, E. & Gächter, S. 2002 Altruistic punishment in humans. *Nature* **415**, 137–140. (doi:10.1038/415137a)

Fehr, E. & Schmidt, K. M. 1999 A theory of fairness, competition and cooperation. *Q. J. Econ.* **114**, 817–868. (doi:10.1162/003355399556151)

Fowler, J. H. 2005 Altruistic punishment and the origin of cooperation. *Proc. Natl Acad. Sci. USA* **102**, 7047–7049. (doi:10.1073/pnas.0500938102)

Fowler, J. H., Johnson, T. & Smirnov, O. 2005 Egalitarian motive and altruistic punishment. *Nature* **433**, E1. (doi:10.1038/nature03256)

Gächter, S. & Herrmann, B. 2006 The limits of self-governance in the presence of spite: experimental evidence from urban and rural Russia. CeDEx discussion paper no. 2006-13.

Gürerk, Ö., Irlenbusch, B. & Rockenbach, B. 2006 The competitive advantage of sanctioning institutions. *Science* **312**, 108–111. (doi:10.1126/science.1123633)

Hagen, E. & Hammerstein, P. 2006 Game theory and human evolution: a critique of some recent interpretations of experimental games. *Theor. Popul. Biol.* **69**, 339–348. (doi:10.1016/j.tpb.2005.09.005)

Henrich, J. *et al.* 2006 Costly punishment across human societies. *Science* **312**, 1767–1770. (doi:10.1126/science.1127333)

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V. & Fehr, E. 2006 Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science* **314**, 829–832. (doi:10.1126/science.1129156)

Kosfeld, M. & Riedl, A. 2007 Order without law? Experimental evidence on voluntary cooperation and sanctioning. *KritV* **90**, 140–155.

Lehmann, L. & Keller, L. 2006 The evolution of cooperation and altruism—a general framework and a classification of models. *J. Evol. Biol.* **19**, 1365–1376. (doi:10.1111/j.1420-9101.2006.01119.x)

Levine, D. K. 1998 Modeling altruism and spitefulness in experiments. *Rev. Econ. Dyn.* **1**, 593–622. (doi:10.1006/redy.1998.0023)

Masclet, D. 2003 Ostracism in work teams: a public good experiment. *Int. J. Manpower* **24**, 867–887. (doi:10.1108/01437720310502177)

Masclet, D., Noussair, C., Tucker, S. & Villeval, M.-C. 2003 Monetary and nonmonetary punishment in the voluntary contributions mechanism. *Am. Econ. Rev.* **93**, 366–380.

Maynard Smith, J. & Harper, D. 2003 *Animal signals*. Oxford, UK: Oxford University Press.

Nikiforakis, N. 2008 Punishment and counter-punishment in public good games: can we really govern ourselves? *J. Pub. Econ.* **92**, 91–112. (doi:10.1016/j.jpubeco.2007.04.008)

Nikiforakis, N. & Normann, H.-T. In press. A comparative statics analysis of punishment in public good experiments. *Exp. Econ.*

Nowak, M. A. & Sigmund, K. 2005 Evolution of indirect reciprocity. *Nature* **437**, 1291–1298. (doi:10.1038/nature04131)

Panchanathan, K. & Boyd, R. 2004 Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature* **432**, 499–502. (doi:10.1038/nature02978)

Rabin, M. 1993 Incorporating fairness into game theory and economics. *Am. Econ. Rev.* **83**, 1281–1302.

Resnick, P. & Zeckhauser, R. 2002 Trust among strangers in internet auctions: empirical analysis of eBay's reputation system. In *The economics of the Internet and e-commerce*, vol. 11 (ed. M. R. Baye). Advances in applied microeconomics, pp. 127–157. Amsterdam, The Netherlands: Elsevier Science.

Rockenbach, B. & Milinski, M. 2006 The efficient interaction of indirect reciprocity and costly punishment. *Nature* **444**, 718–723. (doi:10.1038/nature05229)

Sachs, J. L., Mueller, U. G., Wilcox, T. P. & Bull, J. J. 2004 The evolution of cooperation. *Q. Rev. Biol.* **79**, 135–160. (doi:10.1086/383541)

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E. & Cohen, J. D. 2003 The neural basis of economic decision-making in the ultimatum game. *Science* **300**, 1755–1758. (doi:10.1126/science.1082976)

Sigmund, K., Hauert, C. & Nowak, M. A. 2001 Reward and punishment. *Proc. Natl Acad. Sci. USA* **98**, 10 757–10 762. (doi:10.1073/pnas.161155698)

Sugden, R. 1986 *The economics of rights, cooperation and welfare*. Oxford, UK: Blackwell.

Trivers, R. 1971 The evolution of reciprocal altruism. *Q. Rev. Biol.* **46**, 35–57. (doi:10.1086/406755)

Ule, A. 2005 *Exclusion and cooperation in networks*. Amsterdam, The Netherlands: Tinbergen Institute.

Walker, J. & Halloran, M. 2004 Rewards and sanctions and the provision of public goods in one-shot settings. *Exp. Econ.* **7**, 235–247. (doi:10.1023/B:EXEC.0000040559.08652.51)

West, S. A., Pen, I. & Griffin, A. S. 2002 Cooperation and competition between relatives. *Science* **296**, 72–75. (doi:10.1126/science.1065507)